A Review of Stochastic Gradient Descent

Ravi Raghavan & Lakshitha Ramanayake

March 2024

1 Convex Optimization Basic Concepts

1.1 Smooth Functions and Convexity

When discussing convergence, we will first discuss the underlying theory of Smooth Functions and Convexity.

1.1.1 Differentiability

Definition 1 (Jacobian). Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be differentiable, and $x \in$ **int dom**f. Then, we note $Df(x)$ the derivative or **Jacobian** of f at x, which is the matrix defined by its first partial derivatives:

$$[Df(x)]_{ij} = \frac{\partial f_i}{\partial x_j}(x), \;\; for \;\; i = 1, \ldots, m, j = 1, \ldots, n \tag{1}$$

Remark 2 (Gradient). If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable, then $Df(x) \in \mathbb{R}^{1 \times d}$ is a row vector, whose transpose is called the **gradient** of f at x:

$$\nabla f(x) = Df(x)^T \in \mathbb{R}^{d \times 1} \tag{2}$$

Definition 3 (Hessian). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable, and $x \in$ **int dom**f. Then we note $\nabla^2 f(x)$ the **Hessian** of f at x, which is the matrix defined by its second-order partial derivatives:

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x), \;\; for \;\; i, j = 1, \ldots, n \tag{3}$$

Consequently $\nabla^2 f(x)$ is a $n \times n$ matrix.

Definition 4 (Lipschitz). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and $L > 0$. We say that F is L-**Lipschitz** if

$$\forall x, y \in \textbf{dom} f, \;\; ||f(y) - f(x)|| \leq L||y - x|| \tag{4}$$

1.1.2 Convexity

Definition 5 (Jensen's Inequality). A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if **dom** f is a convex set and if for all $x, y \in dom f$, and θ with $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y) \tag{5}$$

Definition 6 (First Order Condition of Convexity). A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if and only if **dom** f is a convex set and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{6}$$

holds for all $x, y \in dom f$

Definition 7 (Second Order Condition of Convexity). A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if and only if **dom** f is a convex set and its Hessian is positive semi-definite: for all $x \in dom f$,

$$\nabla^2 f(x) \succeq 0 \tag{7}$$

1.1.3 Strong Convexity

Definition 8 (Jensen's Inequality for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if **dom** f is a convex set and if for all $x, y \in domf$, and θ with $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{p}{2}\theta(1-\theta)||x-y||_2^2 \tag{8}$$

Definition 9 (First Order Condition for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if and only if **dom** f is a convex set and

$$f(y) \geq f(x) + \nabla(f(x))^T(y-x) + \frac{p}{2}||y-x||_2^2 \tag{9}$$

holds for all $x, y \in domf$

Definition 10 (Second Order Condition for Strong Convexity). A function $f : \mathbb{R}^n \to \mathbb{R}$ is *p-strongly convex* if and only if **dom** f is a convex set and

$$\nabla^2 f(x) \succeq pI \tag{10}$$

1.1.4 Smoothness

Definition 11 (L-Smooth Functions). A function $f : \mathbb{R}^n \to \mathbb{R}$ and $L > 0$ is L-Smooth if it is differentiable and if $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is $L-$Lipschitz:

$$\forall x, y \in \mathbb{R}^n, \;\; ||\nabla f(y) - \nabla f(x)|| \leq L||y-x|| \tag{11}$$

As a subsequent node, $L-$Smooth functions have a quadratic upper bound:

$$\forall x, y \in \mathbb{R}^n, f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}||y-x||^2 \tag{12}$$

Lemma 12: If f is $L-$smooth and $\gamma > 0$ then,

$$\forall x, y \in \mathbb{R}^n, \;\; f(x - \gamma\nabla f(x)) - f(x) \leq -\gamma(1 - \frac{\gamma L}{2})||\nabla f(x)||^2 \tag{13}$$

Proof:

According to Definition 11, when a function is $L-$Smooth, the following holds true:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x-y|| \tag{14}$$

Let $g(t) = f(x + t(y-x))$. Based on the Fundamental Theorem of Calculus

$$f(y) - f(x) = \int_0^1 g'(t)\, dt \tag{15}$$

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x))^T(y-x)\, dt \tag{16}$$

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y-x)), y-x \rangle\, dt \tag{17}$$

$$f(y) - f(x) = \langle \nabla f(x), y-x \rangle + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y-x \rangle\, dt \tag{18}$$

Applying the Cauchy Schwarz Inequality:

$$f(y) - f(x) \leq \langle \nabla f(x), y-x \rangle + \int_0^1 ||\nabla f(x + t(y-x)) - \nabla f(x)|| \;\; ||y-x||\, dt \tag{19}$$

Now let's apply the definition of $L-$Smoothness inside the integral.

$$f(y) - f(x) \leq \langle \nabla f(x), y-x \rangle + \int_0^1 tL||y-x||^2\, dt \tag{20}$$

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}||y - x||^2 \tag{21}$$

Let's insert $y = x - \gamma \nabla f(x)$ into the above equation

$$f(x - \gamma \nabla f(x)) - f(x) \leq \langle \nabla f(x), x - \gamma \nabla f(x) - x \rangle + \frac{L}{2}||x - \gamma \nabla f(x) - x||^2 \tag{22}$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq \langle \nabla f(x), -\gamma \nabla f(x) \rangle + \frac{L}{2}|| - \gamma \nabla f(x)||^2 \tag{23}$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq -\langle \nabla f(x), \gamma \nabla f(x) \rangle + \frac{L\gamma^2}{2}||\nabla f(x)||^2 \tag{24}$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq -\gamma||\nabla f(x)||^2 + \frac{L\gamma^2}{2}||\nabla f(x)||^2 \tag{25}$$

$$f(x - \gamma \nabla f(x)) - f(x) \leq (-\gamma + \frac{L\gamma^2}{2})||\nabla f(x)||^2 \tag{26}$$

If we assume that $\inf f > -\infty$ and if we set $\gamma = \frac{1}{L}$, we can see that:

$$\inf f - f(x) \leq f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq (-\frac{1}{2L})||\nabla f(x)||^2 \tag{27}$$

$$(\frac{1}{2L})||\nabla f(x)||^2 \leq f(x) - \inf f \tag{28}$$

1.1.5 Smoothness and Convexity

Lemma 13: If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and $L-$smooth, then $\forall x, y \in \mathbb{R}^d$, we have that:

$$\frac{1}{2L}||\nabla f(y) - \nabla f(x)||^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \tag{29}$$

Proof: $f(x) - f(y) = f(x) - f(z) + f(z) - f(y)$

Due to the first order condition of convexity

$$f(z) \geq f(x) + \nabla f(x)^T(z - x) \tag{30}$$

$$f(x) - f(z) \leq -\nabla f(x)^T(z - x) = \nabla f(x)^T(x - z) \tag{31}$$

Since L-Smooth functions have a Quadratic Upper Bound

$$f(z) \leq f(y) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}||z - y||^2 \tag{32}$$

$$f(z) - f(y) \leq \langle \nabla f(y), z - y \rangle + \frac{L}{2}||z - y||^2 \tag{33}$$

$$f(x) - f(y) = f(x) - f(z) + f(z) - f(y) \leq \nabla f(x)^T(x - z) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}||z - y||^2 \tag{34}$$

To minimize the Right Hand Side with respect to z, let's first compute the gradient of the Right Hand Side with respect to z

$$-\nabla f(x) + \nabla f(y) + \frac{L}{2}(2z - 2y) \tag{35}$$

Setting this to zero gives us: $z - y = \frac{1}{L}(\nabla f(x) - \nabla f(y))$, which means that: $z = y - \frac{1}{L}(\nabla f(y) - \nabla f(x))$

Let's substitute this value for z back into the Right Hand Side of Equation (34).

$$\nabla f(x)^T(x-(y-\frac{1}{L}(\nabla f(y)-\nabla f(x))))+\langle \nabla f(y), y-\frac{1}{L}(\nabla f(y)-\nabla f(x))-y\rangle+\frac{L}{2}||y-\frac{1}{L}(\nabla f(y)-\nabla f(x))-y||^2 \tag{36}$$

$$\nabla f(x)^T(x-(y-\frac{1}{L}(\nabla f(y)-\nabla f(x))))-\langle \nabla f(y), \frac{1}{L}(\nabla f(y)-\nabla f(x))\rangle+\frac{L}{2}||\frac{1}{L}(\nabla f(y)-\nabla f(x))||^2 \tag{37}$$

$$\nabla f(x)^T(x-y)+\frac{1}{L}\nabla f(x)^T(\nabla f(y)-\nabla f(x))-\langle \nabla f(y), \frac{1}{L}(\nabla f(y)-\nabla f(x))\rangle+\frac{1}{2L}||\nabla f(y)-\nabla f(x)||^2 \tag{38}$$

$$\nabla f(x)^T(x-y)+\frac{1}{L}\langle \nabla f(x)-\nabla f(y), \nabla f(y)-\nabla f(x)\rangle+\frac{1}{2L}||\nabla f(y)-\nabla f(x)||^2 \tag{39}$$

$$\nabla f(x)^T(x-y)-\frac{1}{L}\langle \nabla f(y)-\nabla f(x), \nabla f(y)-\nabla f(x)\rangle+\frac{1}{2L}||\nabla f(y)-\nabla f(x)||^2 \tag{40}$$

$$\nabla f(x)^T(x-y)-\frac{1}{L}||\nabla f(y)-\nabla f(x)||^2+\frac{1}{2L}||\nabla f(y)-\nabla f(x)||^2 \tag{41}$$

$$\nabla f(x)^T(x-y)-\frac{1}{2L}||\nabla f(y)-\nabla f(x)||^2 \tag{42}$$

We chose the optimal value of z to minimize the Right Hand Side. Hence, it is clear that

$$f(x)-f(y)\leq \nabla f(x)^T(x-y)-\frac{1}{2L}||\nabla f(y)-\nabla f(x)||^2 \tag{43}$$

$$\frac{1}{2L}||\nabla f(y)-\nabla f(x)||^2\leq f(y)-f(x)-\langle \nabla f(x), y-x\rangle \tag{44}$$

1.2 Gradient Descent

Gradient Descent Algorithm. Let $x^{(0)} \in \mathbf{dom} f$, and let $\gamma > 0$ be a step size. The **Gradient Descent (GD)** algorithm defines a sequence $(x^{(t)})_{t\in\mathbb{N}}$ satisfying

$$x^{(t+1)} = x^{(t)} - \gamma \nabla f(x^{(t)}) \tag{45}$$

1.3 Function Definitions

Sum of Functions. Let's say that we have a function $f : \mathbb{R}^n \to \mathbb{R}$ which can be expressed as:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) \tag{46}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$. We can say that f is a Sum of Functions

Sum of Convex Functions. Let's say that we have a function $f : \mathbb{R}^n \to \mathbb{R}$ which can be expressed as:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) \tag{47}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ and f_i is convex. We can say that f is a Sum of Convex Functions

Sum of L-Smooth Functions. Let's say that we have a function $f : \mathbb{R}^n \to \mathbb{R}$ which can be expressed as:

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) \tag{48}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ and f_i is L_i smooth. We can say that f is a Sum of L-Smooth Functions. Let $L_{max} = \max_{\{} 1, \ldots, m\}\{L_i\}$

1.4 Stochastic Gradient Descent

Stochastic Gradient Descent Algorithm. Let's say that we are minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$ which is a sum of functions. It is assumed that $\arg\min f \neq \emptyset$ and that f_i is unbounded below.
Let $x^{(0)} \in \mathbf{dom} f$, and let $\gamma_t > 0$ be a sequence of step sizes. The **Stochastic Gradient Descent (GD)** algorithm defines a sequence $(x^{(t)})_{t\in\mathbb{N}}$ satisfying

$$i_t \in \{1, \ldots, m\} \tag{49}$$

$$x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{i_t}(x^{(t)}) \tag{50}$$

Note i_t is sampled with probability $\frac{1}{m}$ and $E[\nabla f_{i_t}(x)] = \frac{1}{m}\sum_{i=1}^m \nabla f_i(x) = \nabla f(x)$

1.5 Minibatch Stochastic Gradient Descent

Minibatch Stochastic Gradient Descent Algorithm. Let's say that we are minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$ which is a sum of functions. It is assumed that $\arg\min f \neq \emptyset$ and that f_i is unbounded below.
Let $x^{(0)} \in \mathbf{dom} f$, let $b \in [1, m]$ be the batch size, and let $\gamma_t > 0$ be a sequence of step sizes. The **Minibatch Stochastic Gradient Descent (Minibatch SGD)** algorithm defines a sequence $(x^{(t)})_{t\in\mathbb{N}}$ satisfying

$$B_t \subset \{1, \ldots, m\} \tag{51}$$

$$x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{B_t}(x^{(t)}) \tag{52}$$

Note B_t is sampled uniformly among all sets of size b. This means that given a batch of size b, it has a probability of $\frac{1}{\binom{m}{b}}$ of being selected

$$\nabla f_{B_t}(x^{(t)}) = \frac{1}{|B|}\sum_{i\in B} \nabla f_i(x^{(t)}) \tag{53}$$

Observation: $\nabla f_i(x^{(t)})$ will be used when computing mini batch gradients(i.e. $\nabla f_{B_t}(x^{(t)})$) in exactly $\binom{m-1}{b-1}$ mini batches.
Property: $\binom{m}{b} = \binom{m-1}{b-1} \cdot \frac{m}{b}$

$$\mathbb{E}[\nabla f_{B_t}(x^{(t)})] = \frac{1}{\binom{m}{b}}\binom{m-1}{b-1}\sum_{i=1}^m \frac{1}{b}\nabla f_i(x^{(t)}) = \frac{1}{m}\sum_{i=1}^m \nabla f_i(x^{(t)}) = \nabla f(x^{(t)}) \tag{54}$$

1.6 Expected Smoothness and Variance [SGD]

1.6.1 Expected Smoothness

Lemma 14: Let's say that we have a function $f : \mathbb{R}^n \to \mathbb{R}$ which is a sum of Convex functions and a sum of L-Smooth functions, we can state the following:

$$\forall x, y \in \mathbf{dom} f, \;\; \frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(y) - \nabla f_i(x)||^2] \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \tag{55}$$

Proof: Lemma 13 and $L_i \leq L_{max}$ tells us that:

$$\frac{1}{2L_{max}}||\nabla f_i(y) - \nabla f_i(x)||^2 \leq \frac{1}{2L_i}||\nabla f_i(y) - \nabla f_i(x)||^2 \leq f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle \tag{56}$$

$$\frac{1}{2L_{max}}||\nabla f_i(y) - \nabla f_i(x)||^2 \leq f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle \tag{57}$$

We are now prepared to take expectation.

$$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(y) - \nabla f_i(x)||^2] \leq \mathbb{E}[(f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle)] \tag{58}$$

Applying Linearity of Expectation on the Right Hand Side gives us:

$$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(y) - \nabla f_i(x)||^2] \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \tag{59}$$

Lemma 15: When $x = x^*$, where $x^* \in \arg\min f$, and $y = x$, following Lemma 14, we can see that

$$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle \tag{60}$$

Since $\nabla f(x^*) = 0$

$$\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq f(x) - \inf f \tag{61}$$

1.6.2 Variance

Definition 16 (Interpolation). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of Functions. **Interpolation** holds if there exists a common $x^* \in \mathbb{R}^n$ such that $f_i(x^*) = inf f_i, \forall i = 1, \dots m$.

Lemma 17. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of Functions. If interpolation holds at $x^* \in \mathbb{R}^n$, then $x^* \in \arg\min f$

Proof. Since we know that interpolation holds at x^*, we know that $f_i(x^*) = inf f_i, \forall i = 1, \dots m$

We know that the following holds true for all $x \in dom f$

$$f(x^*) = \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = \frac{1}{m}\sum_{i=1}^{m} inf f_i \leq \frac{1}{m}\sum_{i=1}^{m} f_i(x) = f(x) \tag{62}$$

Definition 18(Function Noise). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of Functions. The **Function Noise**, Δ_f^*, is defined as:

$$\Delta_f^* = \inf f - \frac{1}{m}\sum_{i=1}^{m} \inf f_i \tag{63}$$

Lemma 19. Given the previous definition, we can state the following

$$\Delta_f^* \geq 0 \tag{64}$$

Interpolation Holds if and only if $\Delta_f^* = 0$

Proof. Let $x^* \in \arg\min f$

$$\Delta_f^* = \inf f - \frac{1}{m}\sum_{i=1}^{m} \inf f_i = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} \inf f_i \geq f(x^*) - \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = f(x^*) - f(x^*) = 0 \tag{65}$$

Let's now prove that Interpolation Holds if and only if $\Delta_f^* = 0$.
Proof of First Direction(i.e. If Interpolation, then $\Delta_f^* = 0$)
Due to interpolation, $\inf f_i = f_i(x^*)$. Hence,

$$\Delta_f^* = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} inf f_i = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = f(x^*) - f(x^*) = 0 \tag{66}$$

Proof of second direction(i.e. If $\Delta_f^* = 0$, then Interpolation holds)

$$\Delta_f^* = f(x^*) - \frac{1}{m}\sum_{i=1}^{m} \inf f_i \geq f(x^*) - \frac{1}{m}\sum_{i=1}^{m} f_i(x^*) = f(x^*) - f(x^*) = 0 \tag{67}$$

Subsequently,

$$\sum_{i=1}^{m} \inf f_i = \sum_{i=1}^{m} f_i(x^*) \tag{68}$$

$$\sum_{i=1}^{m} (f_i(x^*) - \inf f_i) = 0 \tag{69}$$

Conclusion: $f_i(x^*) = \inf f_i$, meaning Interpolation Holds!

Definition 20. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions. **Gradient Noise**, σ_f^*, is defined as such:

$$\sigma_f^* = \inf_{x^* \in \arg\min f} V[\nabla f_i(x^*)] \tag{70}$$

where $V[X] = E[||X - E[X]||^2]$

Lemma 21. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions

$$\sigma_f^* = V[\nabla f_i(x^*)], \forall x^* \in \arg\min f \tag{71}$$

Proof. Let $x_1, x_2 \in \arg\min f$.
Based on Lemma 15,

$$\frac{1}{2L_{max}} \mathbb{E}[||\nabla f_i(x_1) - \nabla f_i(x_2)||^2] \leq f(x_1) - \inf f = \inf f - \inf f = 0 \tag{72}$$

Subsequently, since $\mathbb{E}[||\nabla f_i(x_1) - \nabla f_i(x_2)||^2] \geq 0$, for the above to be true, $\mathbb{E}[||\nabla f_i(x_1) - \nabla f_i(x_2)||^2] = 0$ and $||\nabla f_i(x_1) - \nabla f_i(x_2)||^2 = 0$. This means that $f_i(x_1) = f_i(x_2)$ and, subsequently, $V[\nabla f_i(x_1)] = V[\nabla f_i(x_2)]$
Lemma 22. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions

1. $\sigma_f^* \leq 2L_{max}\Delta_f^*$
2. If each f_i is p strongly convex, then $2p\Delta_f^* \leq \sigma_f^*$

Proof: Let $x^* \in \arg\min f$.
Based on Lemma 12, $||\nabla f(x)||^2 \leq 2L * (f(x) - \inf f)$.
Since each f_i is $L-$Smooth,

$$||\nabla f_i(x^*)||^2 \leq 2L_i * (f_i(x^*) - \inf f_i) \leq 2L_{max} * (f_i(x^*) - \inf f_i) \tag{73}$$

Taking Expectations on both sides of inequality:

$$\mathbb{E}[||\nabla f_i(x^*)||^2] \leq \mathbb{E}[2L_{max} * (f_i(x^*) - \inf f_i)] \tag{74}$$

$$\mathbb{E}[||\nabla f_i(x^*)||^2] \leq 2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)] \tag{75}$$

Since $\nabla f(x^*) = 0$

$$\mathbb{E}[||\nabla f_i(x^*)||^2] = \mathbb{E}[||\nabla f_i(x^*) - \nabla f(x^*)||^2] = \mathbb{E}[||\nabla f_i(x^*) - \mathbb{E}[\nabla f_i(x^*)]||^2] = \mathbb{V}[\nabla f_i(x^*)] \geq \sigma_f^* \tag{76}$$

Due to Linearity of Expectation,

$$2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)] = 2L_{max}(\mathbb{E}[(f_i(x^*))] - \mathbb{E}[(\inf f_i)]) = 2L_{max}(f(x^*) - \frac{1}{n}\sum_{i=1}^{n} \inf f_i) \tag{77}$$

$$= 2L_{max}(\inf f - \frac{1}{n}\sum_{i=1}^{n} \inf f_i) = 2L_{max}\Delta_f^* \tag{78}$$

Part (2) of Proof: Show that, if each f_i is p strongly convex, then $2p\Delta_f^* \leq \sigma_f^*$

Due to strong convexity, when a function f is p strongly convex and $x^* \in \arg\min f$:

$$f(y) \geq f(x) + \nabla(f(x))^T(y - x) + \frac{p}{2}||y - x||_2^2 \tag{79}$$

Set $y = x^*$ and $x = x$

$$f(x^*) \geq f(x) + \nabla(f(x))^T(x^* - x) + \frac{p}{2}||x^* - x||_2^2 \tag{80}$$

After some mathematical manipulations,

$$f(x)-f(x^*) \leq \nabla(f(x))^T(x-x^*)-\frac{p}{2}||x^*-x||_2^2 = \frac{-1}{2}||\sqrt{p}(x-x^*)-\frac{1}{\sqrt{p}}\nabla f(x)||^2+\frac{1}{2p}||\nabla f(x)||^2 \leq \frac{1}{2p}||\nabla f(x)||^2 \tag{81}$$

Hence,

$$f_i(x) - \inf f_i \leq \frac{1}{2p}||\nabla f_i(x)||^2 \tag{82}$$

$$f_i(x^*) - \inf f_i \leq \frac{1}{2p}||\nabla f_i(x^*)||^2 \tag{83}$$

Take Expectation over this inequality:

$$\mathbb{E}[f_i(x^*) - \inf f_i] \leq \frac{1}{2p}\mathbb{E}[||\nabla f_i(x^*)||^2] \tag{84}$$

Applying Linearity of Expectation:

$$\mathbb{E}[f_i(x^*)] - \mathbb{E}[\inf f_i] \leq \frac{1}{2p}\mathbb{E}[||\nabla f_i(x^*)||^2] \tag{85}$$

Since $\nabla f(x^*) = 0$, $\mathbb{E}[||\nabla f_i(x^*)||^2] = \mathbb{E}[||\nabla f_i(x^*) - \nabla f(x^*)||^2] = \mathbb{E}[||\nabla f_i(x^*) - \mathbb{E}[\nabla f_i(x^*)]||^2] = \mathbb{V}[\nabla f_i(x^*)]$

$$\inf f - \frac{1}{n}\sum_{i=1}^{n} \inf f_i \leq \frac{1}{2p}\mathbb{V}[\nabla f_i(x^*)] \tag{86}$$

$$\Delta_f^* \leq \frac{1}{2p}\mathbb{V}[\nabla f_i(x^*)] \tag{87}$$

Due to convexity, $\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$

$$\Delta_f^* \leq \frac{1}{2p}\sigma_f^* \tag{88}$$

$$2p\Delta_f^* \leq \sigma_f^* \tag{89}$$

Lemma 23. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions

$$\forall x \in \mathbb{R}^d, \;\; \mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}\Delta_f^* \tag{90}$$

Proof: As per Lemma 12,

$$\frac{1}{2L}||\nabla f(x)||^2 \leq f(x) - \inf f \tag{91}$$

Hence,

$$\frac{1}{2L_i}||\nabla f_i(x)||^2 \leq f_i(x) - \inf f_i \tag{92}$$

$$||\nabla f_i(x)||^2 \leq 2L_i(f_i(x)-\inf f_i) \leq 2L_{max}(f_i(x)-\inf f_i) = 2L_{max}(f_i(x)-f_i(x^*))+2L_{max}(f_i(x^*)-\inf f_i) \tag{93}$$

$$||\nabla f_i(x)||^2 \leq 2L_{max}(f_i(x) - f_i(x^*)) + 2L_{max}(f_i(x^*) - \inf f_i) \tag{94}$$

Taking Expectation over both sides of the inequality:

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}\mathbb{E}[(f_i(x) - f_i(x^*))] + 2L_{max}\mathbb{E}[(f_i(x^*) - \inf f_i)] \tag{95}$$

According to Linearity of Expectation:

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(\mathbb{E}[(f_i(x))] - \mathbb{E}[(f_i(x^*))]) + 2L_{max}(\mathbb{E}[(f_i(x^*))] - \mathbb{E}[(\inf f_i)]) \tag{96}$$

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}(\inf f - E[\inf f_i]) \tag{97}$$

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}(\inf f - \frac{1}{n}\sum_{i=1}^{n} \inf f_i) \tag{98}$$

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2L_{max}(f(x) - \inf f) + 2L_{max}\Delta_f^* \tag{99}$$

Lemma 24. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions

$$\forall x \in \mathbb{R}^d,\ \ \mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^* \tag{100}$$

Proof: Let $x^* \in \arg\min f$

$$||\nabla f_i(x)||^2 = ||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2 \tag{101}$$

Based on the triangular inequality for norms, $||x + y|| \leq ||x|| + ||y||$

$$||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2 \leq (||\nabla f_i(x) - \nabla f_i(x^*)|| + ||f_i(x^*)||)^2 \tag{102}$$

$$||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2 \leq ||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x) - \nabla f_i(x^*)||||\nabla f_i(x^*)|| + ||\nabla f_i(x^*)||^2 \tag{103}$$

Known Fact: $||\nabla f_i(x) - \nabla f_i(x^*) - f_i(x^*)||^2 \geq 0$

$$||\nabla f_i(x) - \nabla f_i(x^*)||^2 - 2||\nabla f_i(x) - \nabla f_i(x^*)||||\nabla f_i(x^*)|| + ||\nabla f_i(x^*)||^2 \geq 0 \tag{104}$$

$$||\nabla f_i(x) - \nabla f_i(x^*)||^2 + ||\nabla f_i(x^*)||^2 \geq 2||\nabla f_i(x) - \nabla f_i(x^*)||||\nabla f_i(x^*)|| \tag{105}$$

Using Inequality (105) in Inequality (103):

$$||\nabla f_i(x) - \nabla f_i(x^*) + f_i(x^*)||^2 \leq 2||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x^*)||^2 \tag{106}$$

$$||\nabla f_i(x)||^2 \leq 2||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x^*)||^2 \tag{107}$$

Take the expectation over this inequality

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq \mathbb{E}[2||\nabla f_i(x) - \nabla f_i(x^*)||^2 + 2||\nabla f_i(x^*)||^2] \tag{108}$$

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq 2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2 + ||\nabla f_i(x^*)||^2] \tag{109}$$

According to Lemma 15, $\frac{1}{2L_{max}}\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq f(x) - \inf f$ and $\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq 2L_{max}(f(x) - \inf f)$

$$2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq 4L_{max}(f(x) - \inf f) \tag{110}$$

Since f is a Sum of $L-$Smooth and Convex Functions,

$$\sigma_f^* = V[\nabla f_i(x^*)], \forall x^* \in \arg\min f. \tag{111}$$

$$\sigma_f^* = E[||\nabla f_i(x^*) - E[\nabla f_i(x^*)]||^2], \forall x^* \in \arg\min f. \tag{112}$$

Known Fact: $E[\nabla f_i(x^*)] = \nabla f(x^*) = 0$
Hence, $\sigma_f^* = E[||\nabla f_i(x^*)||^2]$, $2\sigma_f^* = 2E[||\nabla f_i(x^*)||^2]$

Applying Linearity of Expectation on the Right Hand Side of (109)

$$2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2 + ||\nabla f_i(x^*)||^2] = 2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] + 2\mathbb{E}[||\nabla f_i(x^*)||^2] \tag{113}$$

We showed that $2\mathbb{E}[||\nabla f_i(x) - \nabla f_i(x^*)||^2] \leq 4L_{max}(f(x) - \inf f)$ and $2\sigma_f^* = 2E[||\nabla f_i(x^*)||^2])$

Substituting this and (113) into (109) gives:

$$\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^* \tag{114}$$

1.7 Expected Smoothness and Variance [Minibatch SGD]

1.7.1 Expected Smoothness

Definition 25: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions Let $b \in [1, m]$. We can say that f is L_b smooth in expectation if

$$\forall x, y \in \mathbb{R}^n, \frac{1}{2L_b}\mathbb{E}[||\nabla f_B(y) - \nabla f_B(x)||^2] \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \tag{115}$$

If $y = x$ and $x = x^*$, where $x^* \in \arg\min f$, we can see that a function being L_b smooth indicates that:

$$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq f(x) - f(x^*) - \langle \nabla f(x^*), x - x^*\rangle \tag{116}$$

Since $\nabla f(x^*) = 0$

$$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x) - \nabla f_B(x^*)||^2] \leq f(x) - \inf f \tag{117}$$

1.7.2 Variance

Definition 26: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions. **Minibatch Gradient Noise** is defined as such:

$${\sigma_b}^* = \inf_{x^* \in \arg\min f} V[\nabla f_B(x^*)] \tag{118}$$

Lemma 27. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and Convex Functions.

$$\sigma_b^* = V[\nabla f_B(x^*)], \forall x^* \in \arg\min f \tag{119}$$

Proof. Let $x_1, x_2 \in \arg\min f$.

$$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x_1) - \nabla f_B(x_2)||^2] \leq f(x_1) - f(x_2) - \langle \nabla f(x_2), x_1 - x_2\rangle \tag{120}$$

Known Fact: $f(x_2) = \inf f$ and $\nabla f(x_2) = 0$

$$\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x_1) - \nabla f_B(x_2)||^2] \leq f(x_1) - \inf f = 0 \tag{121}$$

Known Fact: $\mathbb{E}[||\nabla f_B(x_1) - \nabla f_B(x_2)||^2] \geq 0$

(121) tell us that $\mathbb{E}[||\nabla f_B(x_1) - \nabla f_B(x_2)||^2] = 0$ and $\nabla f_B(x_1) - \nabla f_B(x_2)$ = 0 and that $\nabla f_B(x_1) = \nabla f_B(x_2)$
Since $f_B(x_1) = f_B(x_2)$, $V[\nabla f_B(x_1)] = V[\nabla f_B(x_2)]$

Lemma 28. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions

$$\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^* \tag{122}$$

Proof: Let $x^* \in \arg\min f$

$$||\nabla f_B(x)||^2 = ||\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)||^2 \tag{123}$$

Triangle Inequality: $||x + y|| \leq ||x|| + ||y||$

$$||\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)||^2 \leq (||\nabla f_B(x) - \nabla f_B(x^*)|| + ||f_B(x^*)||)^2 \tag{124}$$

$$||\nabla f_B(x) - \nabla f_B(x^*) + f_B(x^*)||^2 \leq ||\nabla f_B(x) - \nabla f_B(x^*)||^2 + 2||\nabla f_B(x) - \nabla f_B(x^*)||||\nabla f_B(x^*)|| + ||\nabla f_B(x^*)||^2 \tag{125}$$

Known Fact: $||\nabla f_B(x) - \nabla f_B(x^*) - f_B(x^*)||^2 \geq 0$

$$||\nabla f_B(x)-\nabla f_B(x^*)||^2-2||\nabla f_B(x)-\nabla f_B(x^*)||||\nabla f_B(x^*)||+||\nabla f_B(x^*)||^2\geq 0 \tag{126}$$

$$||\nabla f_B(x)-\nabla f_B(x^*)||^2+||\nabla f_B(x^*)||^2\geq 2||\nabla f_B(x)-\nabla f_B(x^*)||||\nabla f_B(x^*)|| \tag{127}$$

Substituting (127) into (125)

$$||\nabla f_B(x)-\nabla f_B(x^*)+f_B(x^*)||^2\leq 2||\nabla f_B(x)-\nabla f_B(x^*)||^2+2||\nabla f_B(x^*)||^2 \tag{128}$$

$$||\nabla f_B(x)||^2\leq 2||\nabla f_B(x)-\nabla f_B(x^*)||^2+2||\nabla f_B(x^*)||^2 \tag{129}$$

Taking Expectation over both sides of (129):

$$\mathbb{E}[||\nabla f_B(x)||^2]\leq \mathbb{E}[2||\nabla f_B(x)-\nabla f_B(x^*)||^2+2||\nabla f_B(x^*)||^2] \tag{130}$$

$$\mathbb{E}[||\nabla f_B(x)||^2]\leq 2\mathbb{E}[||\nabla f_B(x)-\nabla f_B(x^*)||^2+||\nabla f_B(x^*)||^2] \tag{131}$$

According to Definition 25, $\frac{1}{2L_b}\mathbb{E}[||\nabla f_B(x)-\nabla f_B(x^*)||^2]\leq f(x)-\inf f$ and $\mathbb{E}[||\nabla f_B(x)-\nabla f_B(x^*)||^2]\leq 2L_b(f(x)-\inf f)$

$$2\mathbb{E}[||\nabla f_B(x)-\nabla f_B(x^*)||^2]\leq 4L_b(f(x)-\inf f) \tag{132}$$

Since f is a Sum of $L-$Smooth and Convex Functions, $\sigma_b^*=V[\nabla f_B(x^*)],\forall x^*\in\arg\min f$.

$\sigma_b^*=E[||\nabla f_B(x^*)-E[\nabla f_B(x^*)]||^2],\forall x^*\in\arg\min f$.

Since f is a Sum of Convex Functions, $E[\nabla f_B(x^*)]=\nabla f(x^*)=0$.

In this case, $\sigma_b^*=E[||\nabla f_B(x^*)||^2]$ and $2\sigma_b^*=2E[||\nabla f_B(x^*)||^2]$

Applying Linearity of Expectation to the Right Hand Side of (131):

$$2\mathbb{E}[||\nabla f_B(x)-\nabla f_B(x^*)||^2+||\nabla f_B(x^*)||^2]=2\mathbb{E}[||\nabla f_B(x)-\nabla f_B(x^*)||^2]+2\mathbb{E}[||\nabla f_B(x^*)||^2] \tag{133}$$

We showed that $2\mathbb{E}[||\nabla f_B(x)-\nabla f_B(x^*)||^2]\leq 4L_b(f(x)-\inf f)$ and $2\sigma_b^*=2E[||\nabla f_B(x^*)||^2]$

Subsituting this and (133) into (131)

$$\mathbb{E}[||\nabla f_B(x)||^2]\leq 4L_b(f(x)-\inf f)+2\sigma_b^* \tag{134}$$

2 Varianats on Stochastic Gradient Descent

2.1 Momentum

The key idea behind the momentum method is to incorporate a fraction of the previous update vector into the current update. Let's denote the previous update step as u_{t-1}. Then we can write the current update step u_t according to 135.

$$u^t = \alpha u^{t-1} + \gamma \nabla f(x^{(t)}) \tag{135}$$

where the $alpha \in (0,1)$ is the momentum coefficient, which determines how much of the previous velocity is retained. This coefficient helps in accumulating a direction of persistent descent, smoothing over the updates. With this step, we can write the current iterate as,

$$x^{t+1} = x^t - u^t \tag{136}$$

.

The term u_t is also called the velocity term. The momentum term αu^{t-1} in the velocity term serves as a memory of past gradients:

- If gradients continue pointing in the same direction, the velocity grows in magnitude, allowing for faster convergence.
- If gradients change direction, the velocity's magnitude decreases, which helps mitigate oscillations and overshooting in steep regions of the parameter space.

This approach effectively dampens the oscillations and accelerates convergence towards the minimum of the loss function, particularly in landscapes where the surface curves more steeply in one dimension than in another. To visualize the effect of momentum in optimization let's imagine a ball rolling down a slope. If the slope does not have any turns the ball will keep accumulating velocity till it reaches the bottom. However, if there are turns the ball will slow down to navigate more efficiently.

NEED TO ADD A FIGURE ON MOMENTUM

2.2 Nestrov Accelerated Gradients (NAG)

NAG is also a momentum-based variant of SGD. The main difference between the momentum method and NAG lies in the gradient calculation stage. We have seen that in the momentum method, the update happens at x^t depending on the previous velocity v^{t-1} and the gradient of the function at x^t (135). In NAG the calculation of the gradient is done at a point ahead given by $\nabla f(x^t - \alpha u^{t-1})$.The intuition behind NAS is looking ahead and anticipating, which leads to better solutions. The update rule of NAS can be summarized as mentioned below.

1. Looking Ahead.

$$x^t_{lookahead} = x^t - \alpha v^{t-1} \tag{137}$$

2. Computing the gradient.

$$\nabla f(x^t - \alpha v^{t-1}) \tag{138}$$

3. Taking the gradient step.

$$\begin{aligned} x^{t+1} &= x^t_{lookahead} - \gamma \nabla f(x^t - \alpha v^{t-1}) \\ x^{t+1} &= x^t - \alpha v^{t-1} - \gamma \nabla f(x^t - \alpha v^{t-1}) \end{aligned} \tag{139}$$

This anticipatory step allows NAG to correct its course more responsively than standard Momentum, leading to potentially faster convergence and better handling of the curvature near optimal points. Essentially, NAG adds a level of foresight to updates, which can result in more efficient navigation of complex optimization landscapes.

NEED TO ADD FIGURE ON NESTROV

2.3 Ada Grad

AdaGrad is an adaptive learning rate method that modifies the general approach of gradient descent by allowing each parameter to have its own learning rate. This method addresses a common challenge in training machine learning models, where choosing an appropriate learning rate can be crucial for effective learning. Traditional gradient descent methods use a single learning rate for all parameters, which might not be optimal. Specifically, in a scenario where the users are presented with a dataset with sparse features the methods discussed above take longer to converge to the extremum. This is because the level sets of the problem being elongated balls.

AdaGrad adjusts the learning rate for each parameter based on the history of gradients that have been computed for that parameter. This means that parameters associated with frequently occurring features will have their learning rates decreased, while parameters associated with infrequent features will have their learning rates increased. Such adjustments are beneficial because they make the model less sensitive to the scale of features and more responsive to each feature's specific behavior and importance. This feature-dependent scaling of the learning rate helps in dealing with data sparsity and enhances the convergence properties of the gradient descent optimization, particularly in complex models dealing with high-dimensional data.

Let's define, $f(x)$ to be the stochastic objective function with parameter x, the function evaluation at step t as $f_t(x)$, the gradient of the function with respect to x at step t to be $g_t(s)$. Further take,

$$\mathbf{G}_s = \sum_{t=1}^{s-1} g_t g_t^T \tag{140}$$

Now the update rule for Adagrad can be written as follows.

$$x_{t+1} = x_t - \gamma \mathbf{G}_t^{-\frac{1}{2}} g_t \tag{141}$$

A simplified version of the update rule can be written by only considering the diagonal elements of $\mathbf{G}$.

$$x_{t+1} = x_t - \gamma diag(\mathbf{G}_t)^{-\frac{1}{2}} g_t \tag{142}$$

This simplified version of the update step is computationally efficient when we are dealing with high-dimensional data. Additionally, to avoid the problems arise due to the matrix being singular,in practice a small offset is added to the diagonal elements of the matrix $\mathbf{G}$.

$$x_{t+1} = x_t - \gamma diag(\epsilon \mathbf{I} + \mathbf{G}_t)^{-\frac{1}{2}} g_t \tag{143}$$

Finally, let's look at the expanded version of the update rule.

$$\begin{bmatrix} x_{t+1}^{(1)} \\ x_{t+1}^{(2)} \\ \vdots \\ x_{t+1}^{(m)} \end{bmatrix} = \begin{bmatrix} x_t^{(1)} \\ x_t^{(2)} \\ \vdots \\ x_t^{(m)} \end{bmatrix} - \begin{bmatrix} \frac{\eta}{\sqrt{\varepsilon I + G_t^{(1,1)}}} \\ \frac{\eta}{\sqrt{\varepsilon I + G_t^{(2,2)}}} \\ \vdots \\ \frac{\eta}{\sqrt{\varepsilon I + G_t^{(m,m)}}} \end{bmatrix} \odot \begin{bmatrix} g_t^{(1)} \\ g_t^{(2)} \\ \vdots \\ g_t^{(m)} \end{bmatrix}$$

Where $\odot$ is the Hadamard product between two matrices having the same dimensions. This provides a clear idea of how the per-parameter learning rate works. Here, γ is the paramter which describes the global learning rate. It must also be noted that as G accumulates, the learning rate slows down for each parameter and eventually no progress can be made, causing the algorithm to never reach the exact minima. This is once of the major disadvantages of Adagrad.

2.4 Root Mean Square Propagation (RMS Prop)

We have seen that the monotonically decreasing learning rate of Adagrad leads to a scenario where the learning rate becomes too small too quickly while the descent method can still achieve a reduction of the cost function. RMS prop addresses this issue by introducing a moving window of fixed size over the gradients computed at each step rather than using the full set of gradients. Now the term $G^{(i)}$ for a the coordinate $x^{(i)}$ on the t^{th} iteration can be written as,

$$G_t^{(i)} = \frac{\left(g_{t-w}^{(i)}\right)^2 + \left(g_{t-w+1}^{(i)}\right)^2 + \ldots + \left(g_{t-1}^{(i)}\right)^2}{w} \tag{144}$$

A conceptually equivalent and computationally cheaper way of doing this is to treat 144 as an accumulation of exponentially decaying average of square of gradients. Let, ρ be the decaying factor, then we can write,

$$\mathbb{E}\left[\left(g_t^{(i)}\right)^2\right] = \rho\mathbb{E}\left[\left(g_{t-1}^{(i)}\right)^2\right] + (1-\rho)\left(g_t^{(i)}\right)^2 \tag{145}$$

From 145 we can see that the decay factor causes the older gradient to decay with iteration. This prevents the learning rate from becoming too small too quickly. Now $\left(G_t^i\right)^{-\frac{1}{2}}$ can be seen as,

$$RMS[g_t^{(i)}] = \left(G_{(t)}^i\right)^{-\frac{1}{2}} = \sqrt{\mathbb{E}\left[\left(g_t^{(i)}\right)^2\right]} \tag{146}$$

Finally the update step of RMSprop algorithm,

$$\begin{bmatrix} x_{t+1}^{(1)} \\ x_{t+1}^{(2)} \\ \vdots \\ x_{t+1}^{(m)} \end{bmatrix} = \begin{bmatrix} x_t^{(1)} \\ x_t^{(2)} \\ \vdots \\ x_t^{(m)} \end{bmatrix} - \begin{bmatrix} \frac{\eta}{RMS[g_t^{(1)}]} \\ \frac{\eta}{RMS[g_t^{(2)}]} \\ \vdots \\ \frac{\eta}{RMS[g_t^{(m)}]} \end{bmatrix} \odot \begin{bmatrix} g_t^{(1)} \\ g_t^{(2)} \\ \vdots \\ g_t^{(m)} \end{bmatrix}$$

3 Convergence Behavior of SGD and Minibatch SGD

3.1 SGD Convergence for Convex and Smooth Functions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions. Let the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a sequence of step sizes that satisfy $0 < \gamma_t < \frac{1}{4L_{max}}$.
Denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=0}^{T-1}\gamma_t} + 2\sigma_f^* \frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}$
Proof:
Let $x^* \in \arg\min f$. According to Lemma 21, $\sigma_f^* = \mathbb{V}[\nabla f_i(x^*)]$.
In SGD, the iterates are as follows: $x^{(t+1)} = x^{(t)} - \gamma_t \nabla f_{i_t}(x^{(t)})$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma_t \nabla f_{i_t}(x^{(t-1)}) - x^*||^2 \tag{147}$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma_t \nabla f_{i_t}(x^{(t-1)})||^2 \tag{148}$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t \nabla f_{i_t}(x^{(t-1)})\rangle + ||\gamma_t \nabla f_{i_t}(x^{(t-1)})||^2 \tag{149}$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t \nabla f_{i_t}(x^{(t-1)})\rangle + \gamma_t^2||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{150}$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f_{i_t}(x^{(t-1)})\rangle + \gamma_t^2||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{151}$$

Taking the Expectation conditioned on $x^{(t-1)}$:

$$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{152}$$

Due to the definition of convexity, $f(y) \geq f(x) + \nabla(f(x))^T(y-x)$

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) \tag{153}$$

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) \tag{154}$$

$$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) \tag{155}$$

Substituting (155) into (152)

$$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{156}$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{157}$$

Lemma 24 states that $\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$
Substituting this into (157)

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2(4L_{max}(f(x) - \inf f) + 2\sigma_f^*) \tag{158}$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2 4L_{max}(f(x) - \inf f) + 2\gamma_t^2\sigma_f^* \tag{159}$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 + (2\gamma_t)(2\gamma_t L_{max} - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_f^* \tag{160}$$

Since $\gamma_t < \frac{1}{4L_{max}}$, $2\gamma_t L_{max} - 1 < \frac{-1}{2}$. We also know that $(f(x^{(t-1)}) - f(x^*)) > 0$.

Hence,

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - \gamma_t(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2 \sigma_f^* \tag{161}$$

Taking Expectation over both sides of (161):

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \gamma_t \mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2 \sigma_f^* \tag{162}$$

$$\gamma_t \mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2 \sigma_f^* \tag{163}$$

$$\gamma_t \mathbb{E}(f(x^{(t-1)}) - \inf f) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2 \sigma_f^* \tag{164}$$

Let's build this up recursively

$$\gamma_1 \mathbb{E}(f(x^{(0)}) - \inf f) \leq \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(1)} - x^*||^2 + 2\gamma_1^2 \sigma_f^* \tag{165}$$

$$\gamma_2 \mathbb{E}(f(x^{(1)}) - \inf f) \leq \mathbb{E}||x^{(1)} - x^*||^2 - \mathbb{E}||x^{(2)} - x^*||^2 + 2\gamma_2^2 \sigma_f^* \tag{166}$$

$$\gamma_3 \mathbb{E}(f(x^{(2)}) - \inf f) \leq \mathbb{E}||x^{(2)} - x^*||^2 - \mathbb{E}||x^{(3)} - x^*||^2 + 2\gamma_3^2 \sigma_f^* \tag{167}$$

$$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}(f(x^{(t-1)}) - f(x^*)) = \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(T)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^* \tag{168}$$

Known Fact: $\mathbb{E}||x^{(T)} - x^*||^2 > 0$.

$$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}[(f(x^{(t-1)}) - f(x^*))] = \mathbb{E}[||x^{(0)} - x^*||^2] - \mathbb{E}[||x^{(T)} - x^*||^2] + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^* \leq \mathbb{E}[||x^{(0)} - x^*||^2] + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^* \tag{169}$$

$$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}[(f(x^{(t-1)}) - f(x^*))] \leq \mathbb{E}[||x^{(0)} - x^*||^2] + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^* \tag{170}$$

$$\sum_{t=1}^{T-1} \gamma_t \mathbb{E}[(f(x^{(t-1)}) - f(x^*))] \leq ||x^{(0)} - x^*||^2 + \sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^* \tag{171}$$

Divide both sides of this inequality by $\sum_{t=1}^{T-1} \gamma_t$

$$\mathbb{E}[\sum_{t=1}^{T-1} \frac{\gamma_t}{\sum_{t=1}^{T-1} \gamma_t}(f(x^{(t-1)}) - f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=1}^{T-1} \gamma_t} \tag{172}$$

$$\mathbb{E}[\sum_{t=1}^{T-1} \frac{1}{\sum_{t=1}^{T-1} \gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_f^*}{\sum_{t=1}^{T-1} \gamma_t} \tag{173}$$

Since f is convex, the Generalized Jensen's Inequality means that:

$$f(\bar{x}^T) \leq \frac{1}{\sum_{t=1}^{T-1} \gamma_t} \sum_{t=1}^{T-1} y_t f(x^{(t-1)}) \tag{174}$$

Known Equation:

$$f(x^*) = \frac{1}{\sum_{t=1}^{T-1} \gamma_t} \sum_{t=1}^{T-1} y_t f(x^*) \tag{175}$$

Since $f(x^*) = \inf f$, (174), (175), and the Properties of Linearity Expectation

$$\mathbb{E}[f(\bar{x}^T) - \inf f] = \mathbb{E}[f(\bar{x}^T)] - \mathbb{E}[\inf f] = \mathbb{E}[f(\bar{x}^T)] - \mathbb{E}[f(x^*)] \leq \mathbb{E}[\frac{1}{\sum_{t=1}^{T-1}\gamma_t}\sum_{t=1}^{T-1} y_t f(x^{(t-1)})] - \mathbb{E}[\frac{1}{\sum_{t=1}^{T-1}\gamma_t}\sum_{t=1}^{T-1} y_t f(x^*)] \tag{176}$$

Using (173),

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \mathbb{E}[\sum_{t=1}^{T-1}\frac{1}{\sum_{t=1}^{T-1}\gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*}{\sum_{t=1}^{T-1}\gamma_t} \tag{177}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*}{\sum_{t=1}^{T-1}\gamma_t} \tag{178}$$

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a constant step size $\gamma_t = \gamma \leq \frac{1}{4L_{max}}$.
Denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t = \frac{1}{\gamma T}\gamma\sum_{t=0}^{T-1} x^t = \frac{1}{T}\sum_{t=0}^{T-1} x^t$

Then for every $T \geq 1$, $\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\gamma\sigma_f^*$

Proof. Known Facts: $\sum_{t=0}^{T-1}\gamma_t = \gamma T$ and $\sum_{t=0}^{T-1}\gamma_t^2 = T\gamma^2$.

We have already shown that

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2\sigma_f^*}{\sum_{t=1}^{T-1}\gamma_t} \tag{179}$$

Further Simplification Shows:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + \frac{2\sigma_f^* T\gamma^2}{\gamma T} \tag{180}$$

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\gamma T} + 2\sigma_f^*\gamma \tag{181}$$

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Sum of $L-$ Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a vanishing step size $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ where $\gamma_0 \leq \frac{1}{4L_{max}}$
Denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t$

Then for every $T \geq 1$, $\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_f^*\frac{5\gamma_0\log{(T+1)}}{\sqrt{T}} = \mathbb{O}(\frac{\log{(T+1)}}{\sqrt{T}})$

Proof. We know that our stepsize is decreasing. Hence, we can clearly see that $\gamma_t \leq \gamma_0 \leq \frac{1}{4L_{max}}$ for $t \geq 0$. Hence, we can apply the earlier result that we derived:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1}\gamma_t} + 2\sigma_f^*\frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t} \tag{182}$$

Based on the Sum-Integral Bounds, $\sum_{t=0}^{T-1}\gamma_t = \gamma_0\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \geq \frac{4\gamma_0}{5}\sqrt{T}$ and $\sum_{t=0}^{T-1}\gamma_t^2 = \gamma_0^2\sum_{t=1}^{T}\frac{1}{t} \leq 2\gamma_0^2\log(T+1)$

Substituting it into the earlier inequality, we get:

$$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)} - x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_f^*\frac{5\gamma_0\log{(T+1)}}{\sqrt{T}} = \mathbb{O}(\frac{\log{(T+1)}}{\sqrt{T}}) \tag{183}$$

Using the Sum-Integral Bounds, We will take a brief aside to discuss some theory that will be helpful to prove the theorem.

Let's say we have a function $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ that is decreasing.

It is clear to see that $\int_1^{T+1} f(x)\,dx \leq \sum_{x=1}^{T} f(x) = f(1) + \sum_{x=2}^{T} f(x) \leq f(1) + \int_1^{T} f(x)\,dx$

$\int_1^{T+1} f(x)\,dx \leq \sum_{x=1}^{T} f(x) \leq f(1) + \int_1^{T} f(x)\,dx$

Let's now use the functions $f(x) = \frac{1}{\sqrt{x}}$ and $f(x) = \frac{1}{x}$
Let's start with the function $f(x) = \frac{1}{\sqrt{x}}$

$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx \leq \sum_{x=1}^{T} \frac{1}{\sqrt{x}} \leq 1 + \int_1^{T} \frac{1}{\sqrt{x}}\,dx$

Let's start by working towards the lower bound
$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_1^{T+1} = 2\sqrt{T+1} - 2 = 2(\sqrt{T+1} - 1)$

We know that $\inf_{T\geq 1} \frac{\sqrt{T+1}-1}{\sqrt{T}} = \sqrt{2} - 1 > \frac{2}{5}$

$\int_1^{T+1} \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_1^{T+1} = 2\sqrt{T+1} - 2 = 2(\sqrt{T+1} - 1) \geq \frac{4}{5}\sqrt{T}$

Now let's look at the upper bound:
$1 + \int_1^{T} \frac{1}{\sqrt{x}}\,dx = 1 + 2\sqrt{T+1} - 2 = 2\sqrt{T+1} - 1$
Combining both the Lower and Upper Bounds gives us:

$\frac{4}{5}\sqrt{T} \leq \sum_{x=1}^{T} \frac{1}{\sqrt{x}} \leq 2\sqrt{T+1} - 1$

Now let's analyze the other function $f(x) = \frac{1}{x}$

$\int_1^{T+1} \frac{1}{x}\,dx \leq \sum_{x=1}^{T} \frac{1}{x} \leq 1 + \int_1^{T} \frac{1}{x}\,dx$

Let's start by working towards the lower bound
$\int_1^{T+1} \frac{1}{x}\,dx = [\log(t)]_1^{T+1} = \log(T+1)$

Now let's look at the upper bound:
$1 + \int_1^{T} \frac{1}{x}\,dx = 1 + [\log(t)]_1^{T} = 1 + \log(T) \leq 2\log(T+1)$
Note: We know that $\sup_{T\geq 1} \frac{1+\log(T)}{\log(T+1)} \approx \sqrt{2} < 2$

$\log(T+1) \leq \sum_{x=1}^{T} \frac{1}{x} \leq 2\log(T+1)$

3.2 SGD Convergence for Strongly Convex and Smooth Functions

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. We will also assume that f is p strongly convex. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ and we will assume that we have a constant stepsize satisfying $0 < \gamma < \frac{1}{2L_{max}}$. Then, we can say that for each iteration(i.e. $t \geq 0$)

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1-\gamma p)^t ||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_f^* \tag{184}$$

Proof:
In SGD, the iterates are as follows: $x^{(t+1)} = x^{(t)} - \gamma \nabla f_{i_t}(x^{(t)})$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma \nabla f_{i_t}(x^{(t-1)}) - x^*||^2 \tag{185}$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma \nabla f_{i_t}(x^{(t-1)})||^2 \tag{186}$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{i_t}(x^{(t-1)})\rangle + ||\gamma \nabla f_{i_t}(x^{(t-1)})||^2 \tag{187}$$

$$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma \nabla f_{i_t}(x^{(t-1)})\rangle + \gamma^2||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{188}$$

Take the Expectation conditioned on $x^{(t-1)}$

$$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{189}$$

$f(y) \geq f(x) + \nabla(f(x))^T(y - x) + \frac{p}{2}||y - x||_2^2$ as per Strong Convexity

Hence, $f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2$

$$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)}) + \frac{p}{2}||x^* - x^{(t-1)}||_2^2 \tag{190}$$

$$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2 \tag{191}$$

Substituting this into (189):

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*) + \frac{p}{2}||x^{(t-1)} - x^*||_2^2) + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{192}$$

Simplifying RHS on (192):

$$||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) - p\gamma||x^{(t-1)} - x^*||_2^2 + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{193}$$

$$(1 - p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{194}$$

Putting (192) and (194) together:

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2\mathbb{E}||\nabla f_{i_t}(x^{(t-1)})||^2 \tag{195}$$

As per Lemma 24, $\mathbb{E}[||\nabla f_i(x)||^2] \leq 4L_{max}(f(x) - \inf f) + 2\sigma_f^*$

Substituting Lemma 24 into (195)

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 - 2\gamma(f(x^{(t-1)}) - f(x^*)) + \gamma^2(4L_{max}(f(x^{(t-1)}) - \inf f) + 2\sigma_f^*) \tag{196}$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_{max} - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_f^* \tag{197}$$

Again, taking Expectation on both sides

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_{max} - 1)\mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_f^* \tag{198}$$

Since $\gamma < \frac{1}{2L_{max}}$, $2\gamma L_{max} - 1 < 0$. We also know that $\mathbb{E}(f(x^{(t-1)}) - f(x^*)) > 0$ Hence

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + (2\gamma)(2\gamma L_{max} - 1)\mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma^2\sigma_f^* \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + 2\gamma^2\sigma_f^* \tag{199}$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(t-1)} - x^*||^2 + 2\gamma^2\sigma_f^* \tag{200}$$

Let's build this inequality recursively

$$\mathbb{E}||x^{(1)} - x^*||^2 \leq (1 - p\gamma)||x^{(0)} - x^*||^2 + 2\gamma^2\sigma_f^* \tag{201}$$

$$\mathbb{E}||x^{(2)} - x^*||^2 \leq (1 - p\gamma)\mathbb{E}||x^{(1)} - x^*||^2 + 2\gamma^2\sigma_f^* \tag{202}$$

$$\mathbb{E}||x^{(2)} - x^*||^2 \leq (1 - p\gamma)((1 - p\gamma)||x^{(0)} - x^*||^2 + 2\gamma^2\sigma_f^*) + 2\gamma^2\sigma_f^* \tag{203}$$

We can see that $\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^*$
Let's look at the term $\sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^*$.

$$\sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^* < \sum_{n=0}^{\infty}(1 - p\gamma)^n 2\gamma^2\sigma_f^* = \frac{1}{p\gamma}2\gamma^2\sigma_f^* = \frac{2\gamma\sigma_f^*}{p} \tag{204}$$

Hence, we can see that

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \sum_{n=0}^{t-1}(1 - p\gamma)^n 2\gamma^2\sigma_f^* \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_f^* \tag{205}$$

$$\mathbb{E}||x^{(t)} - x^*||^2 \leq (1 - \gamma p)^t||x^{(0)} - x^*||^2 + \frac{2\gamma}{p}\sigma_f^* \tag{206}$$

3.3 Minibatch SGD Convergence for Convex and Smooth Functions

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the Minibatch SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a sequence of step sizes that satisfy $0 < \gamma_t < \frac{1}{4L_b}$.

Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{t=0}^{T-1}\gamma_t x^t$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=0}^{T-1}\gamma_t} + 2\sigma_b^*\frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}$

Proof:
Let us have $x^* \in \arg\min f$. We have already showed that when f is a Sum of Convex functions, $\sigma_b^* = \mathbb{V}[\nabla f_B(x^*)]$.

We know that, in Minibatch Stochastic Gradient Descent, our iterates progress as such: $x^{(t+1)} = x^{(t)} - \gamma_t\nabla f_{B_t}(x^{(t)})$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - \gamma_t\nabla f_{B_t}(x^{(t-1)}) - x^*||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^* - \gamma_t\nabla f_{B_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t\nabla f_{B_t}(x^{(t-1)})\rangle + ||\gamma_t\nabla f_{B_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\langle x^{(t-1)} - x^*, \gamma_t\nabla f_{B_t}(x^{(t-1)})\rangle + \gamma_t^2||\nabla f_{B_t}(x^{(t-1)})||^2$

$||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f_{B_t}(x^{(t-1)})\rangle + \gamma_t^2||\nabla f_{B_t}(x^{(t-1)})||^2$

Now, let's take the Expectation conditioned on $x^{(t-1)}$

$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Based on the definition of convexity, we know that $f(y) \geq f(x) + \nabla(f(x))^T(y - x)$

This would mean that $f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$

$f(x^*) \geq f(x^{(t-1)}) + \nabla(f(x^{(t-1)}))^T(x^* - x^{(t-1)})$

$\nabla(f(x^{(t-1)}))^T(x^{(t-1)} - x^*) \geq f(x^{(t-1)}) - f(x^*)$

We can substitute this into the earlier equation we derived and get:

$\mathbb{E}||x^{(t)} - x^*||^2 = ||x^{(t-1)} - x^*||^2 - 2\gamma_t\langle x^{(t-1)} - x^*, \nabla f(x^{(t-1)})\rangle + \gamma_t^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Earlier, we proved that, when we have a function that is a sum of $L-$ Smooth functions and that is a sum of convex functions, $\mathbb{E}[||\nabla f_B(x)||^2] \leq 4L_b(f(x) - \inf f) + 2\sigma_b^*$

We can substitute this into the equations we derived:

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2(4L_b(f(x) - \inf f) + 2\sigma_b^*)$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - 2\gamma_t(f(x^{(t-1)}) - f(x^*)) + \gamma_t^2 4L_b(f(x) - \inf f) + 2\gamma_t^2\sigma_b^*$

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 + (2\gamma_t)(2\gamma_t L_b - 1)(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_b^*$

Since $\gamma_t < \frac{1}{4L_b}$, $2\gamma_t L_b - 1 < \frac{-1}{2}$. We also know that $(f(x^{(t-1)}) - f(x^*)) > 0$ Hence

$\mathbb{E}||x^{(t)} - x^*||^2 \leq ||x^{(t-1)} - x^*||^2 - \gamma_t(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_b^*$

Once again, let's take expectation over both sides of this inequality

$\mathbb{E}||x^{(t)} - x^*||^2 \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) + 2\gamma_t^2\sigma_b^*$

$\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2\sigma_b^*$

$\gamma_t\mathbb{E}(f(x^{(t-1)}) - \inf f) \leq \mathbb{E}||x^{(t-1)} - x^*||^2 - \mathbb{E}||x^{(t)} - x^*||^2 + 2\gamma_t^2\sigma_b^*$

Let's build this up recursively:
$\gamma_1\mathbb{E}(f(x^{(0)}) - \inf f) \leq \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(1)} - x^*||^2 + 2\gamma_1^2\sigma_b^*$

$\gamma_2\mathbb{E}(f(x^{(1)}) - \inf f) \leq \mathbb{E}||x^{(1)} - x^*||^2 - \mathbb{E}||x^{(2)} - x^*||^2 + 2\gamma_2^2\sigma_b^*$

$\gamma_3\mathbb{E}(f(x^{(2)}) - \inf f) \leq \mathbb{E}||x^{(2)} - x^*||^2 - \mathbb{E}||x^{(3)} - x^*||^2 + 2\gamma_3^2\sigma_b^*$

$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) = \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(T)} - x^*||^2 + \sum_{t=1}^{T-1}2\gamma_t^2\sigma_b^*$
We know that $\mathbb{E}||x^{(T)} - x^*||^2 > 0$. Hence, we can work with this inequality as such:

$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) = \mathbb{E}||x^{(0)} - x^*||^2 - \mathbb{E}||x^{(T)} - x^*||^2 + \sum_{t=1}^{T-1}2\gamma_t^2\sigma_b^* \leq \mathbb{E}||x^{(0)} - x^*||^2 + + \sum_{t=1}^{T-1}2\gamma_t^2\sigma_b^*$

$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq \mathbb{E}||x^{(0)} - x^*||^2 + \sum_{t=1}^{T-1}2\gamma_t^2\sigma_b^*$

$\sum_{t=1}^{T-1}\gamma_t\mathbb{E}(f(x^{(t-1)}) - f(x^*)) \leq ||x^{(0)} - x^*||^2 + \sum_{t=1}^{T-1}2\gamma_t^2\sigma_b^*$

Let's divide both sides of this inequality by $\sum_{t=1}^{T-1}\gamma_t$

$\mathbb{E}[\sum_{t=1}^{T-1}\frac{\gamma_t}{\sum_{t=1}^{T-1}\gamma_t}(f(x^{(t-1)}) - f(x^*))] \leq \frac{||x^{(0)} - x^*||^2}{\sum_{t=1}^{T-1}\gamma_t} + \frac{\sum_{t=1}^{T-1}2\gamma_t^2\sigma_b^*}{\sum_{t=1}^{T-1}\gamma_t}$

$\mathbb{E}[\sum_{t=1}^{T-1} \frac{1}{\sum_{t=1}^{T-1} \gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$

We know that f is convex. Hence, we can apply the Generalized Jensen's Inequality.
Based on the Generalized Jensen's Inequality, we can see that:

$f(\bar{x}^T) \leq \frac{1}{\sum_{t=1}^{T-1} \gamma_t} \sum_{t=1}^{T-1} y_t f(x^{(t-1)})$
Hence, our inequality becomes

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \mathbb{E}[\sum_{t=1}^{T-1} \frac{1}{\sum_{t=1}^{T-1} \gamma_t}(\gamma_t f(x^{(t-1)}) - \gamma_t f(x^*))] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a constant step size $\gamma_t = \gamma \leq \frac{1}{4L_b}$.

Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t = \frac{1}{\gamma T} \gamma \sum_{t=0}^{T-1} x^t = \frac{1}{T} \sum_{t=0}^{T-1} x^t$

Then for every $T \geq 1$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\gamma T} + 2\gamma\sigma_b^*$

Proof. The proof of this is very simple as we did the majority of heavylifting in the last proof. We know that $\sum_{t=0}^{T-1} \gamma_t = \gamma T$ and $\sum_{t=0}^{T-1} \gamma_t^2 = T\gamma^2$. Let us substitute this into the last theorem and we will proceed from there.

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=1}^{T-1} \gamma_t} + \frac{\sum_{t=1}^{T-1} 2\gamma_t^2 \sigma_b^*}{\sum_{t=1}^{T-1} \gamma_t}$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\gamma T} + \frac{2\sigma_b^* T \gamma^2}{\gamma T}$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\gamma T} + 2\sigma_b^* \gamma$

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. Let us say that the sequence of iterates generated by the SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ with a vanishing step size $\gamma_t = \frac{\gamma_0}{\sqrt{t+1}}$ where $\gamma_0 \leq \frac{1}{4L_b}$

Let us denote $\bar{x}^T = \frac{1}{\sum_{t=0}^{T-1} \gamma_t} \sum_{t=0}^{T-1} \gamma_t x^t$

Then for every $T \geq 1$

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)}-x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_b^* \frac{5\gamma_0 \log{(T+1)}}{\sqrt{T}} = \mathbb{O}(\frac{\log{(T+1)}}{\sqrt{T}})$

Proof. We know that our stepsize is decreasing. Hence, we can clearly see that $\gamma_t \leq \gamma_0 \leq \frac{1}{4L_b}$ for $t \geq 0$. Hence, we can apply the earlier result that we derived:

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{||x^{(0)}-x^*||^2}{\sum_{t=0}^{T-1} \gamma_t} + 2\sigma_b^* \frac{\sum_{t=0}^{T-1} \gamma_t^2}{\sum_{t=0}^{T-1} \gamma_t}$

Based on the Sum-Integral Bounds, we can say that $\sum_{t=0}^{T-1} \gamma_t = \gamma_0 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \geq \frac{4\gamma_0}{5}\sqrt{T}$

$\sum_{t=0}^{T-1} \gamma_t^2 = \gamma_0^2 \sum_{t=1}^{T} \frac{1}{t} \leq 2\gamma_0^2 \log(T+1)$

Substituting it into the expected value, we get:

$\mathbb{E}[f(\bar{x}^T) - \inf f] \leq \frac{5||x^{(0)}-x^*||^2}{4\gamma_0\sqrt{T}} + \sigma_b^* \frac{5\gamma_0 \log{(T+1)}}{\sqrt{T}} = \mathbb{O}(\frac{\log{(T+1)}}{\sqrt{T}})$

3.4 Minibatch SGD Convergence for Strongly Convex and Smooth Functions

Theorem. Let us say that we have a function f that is both a Sum of $L-$Smooth Functions and a Sum of Convex Functions. We will also assume that f is p strongly convex. Let us say that the sequence of iterates generated by the Minibatch SGD Algorithm is $(x^{(t)})_{t\in\mathbb{N}}$ and we will assume that we have a constant stepsize satisfying $0<\gamma<\frac{1}{2L_b}$. Then, we can say that for each iteration(i.e. $t\geq 0$)

$$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-\gamma p)^t||x^{(0)}-x^*||^2+\frac{2\gamma}{p}\sigma_b^* \tag{207}$$

Proof:
We know that, in Minibatch Stochastic Gradient Descent, our iterates progress as such: $x^{(t+1)}=x^{(t)}-\gamma\nabla f_{B_t}(x^{(t)})$

$||x^{(t)}-x^*||^2=||x^{(t-1)}-\gamma\nabla f_{B_t}(x^{(t-1)})-x^*||^2$

$||x^{(t)}-x^*||^2=||x^{(t-1)}-x^*-\gamma\nabla f_{B_t}(x^{(t-1)})||^2$

$||x^{(t)}-x^*||^2=||x^{(t-1)}-x^*||^2-2\langle x^{(t-1)}-x^*,\gamma\nabla f_{B_t}(x^{(t-1)})\rangle+||\gamma\nabla f_{B_t}(x^{(t-1)})||^2$
$||x^{(t)}-x^*||^2=||x^{(t-1)}-x^*||^2-2\langle x^{(t-1)}-x^*,\gamma\nabla f_{B_t}(x^{(t-1)})\rangle+\gamma^2||\nabla f_{B_t}(x^{(t-1)})||^2$
Now, let's take the Expectation conditioned on $x^{(t-1)}$
$\mathbb{E}||x^{(t)}-x^*||^2=||x^{(t-1)}-x^*||^2-2\gamma\langle x^{(t-1)}-x^*,\nabla f(x^{(t-1)})\rangle+\gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Based on the definition of strong convexity, we know that $f(y)\geq f(x)+\nabla(f(x))^T(y-x)+\frac{p}{2}||y-x||_2^2$

This would mean that $f(x^*)\geq f(x^{(t-1)})+\nabla(f(x^{(t-1)}))^T(x^*-x^{(t-1)})+\frac{p}{2}||x^*-x^{(t-1)}||_2^2$

$f(x^*)\geq f(x^{(t-1)})+\nabla(f(x^{(t-1)}))^T(x^*-x^{(t-1)})+\frac{p}{2}||x^*-x^{(t-1)}||_2^2$

$\nabla(f(x^{(t-1)}))^T(x^{(t-1)}-x^*)\geq f(x^{(t-1)})-f(x^*)+\frac{p}{2}||x^{(t-1)}-x^*||_2^2$

We can substitute this into the earlier equation we derived and get:

$\mathbb{E}||x^{(t)}-x^*||^2=||x^{(t-1)}-x^*||^2-2\gamma\langle x^{(t-1)}-x^*,\nabla f(x^{(t-1)})\rangle+\gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2\leq||x^{(t-1)}-x^*||^2-2\gamma(f(x^{(t-1)})-f(x^*)+\frac{p}{2}||x^{(t-1)}-x^*||_2^2)+\gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Let's try to simplify $||x^{(t-1)}-x^*||^2-2\gamma(f(x^{(t-1)})-f(x^*)+\frac{p}{2}||x^{(t-1)}-x^*||_2^2)+\gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

$||x^{(t-1)}-x^*||^2-2\gamma(f(x^{(t-1)})-f(x^*))-p\gamma||x^{(t-1)}-x^*||_2^2+\gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

$(1-p\gamma)||x^{(t-1)}-x^*||^2-2\gamma(f(x^{(t-1)})-f(x^*))+\gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

$\mathbb{E}||x^{(t)}-x^*||^2\leq(1-p\gamma)||x^{(t-1)}-x^*||^2-2\gamma(f(x^{(t-1)})-f(x^*))+\gamma^2\mathbb{E}||\nabla f_{B_t}(x^{(t-1)})||^2$

Earlier, we proved that, when we have a function that is a sum of $L-$ Smooth functions and that is a sum of convex functions, $\mathbb{E}[||\nabla f_B(x)||^2]\leq 4L_b(f(x)-\inf f)+2\sigma_b^*$

We can continue on with our proof as such:

$\mathbb{E}||x^{(t)}-x^*||^2\leq(1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2-2\gamma(f(x^{(t-1)})-f(x^*))+\gamma^2(4L_b(f(x^{(t-1)})-\inf f)+2\sigma_b^*)$

$\mathbb{E}||x^{(t)}-x^*||^2\leq(1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2+(2\gamma)(2\gamma L_b-1)(f(x^{(t-1)})-f(x^*))+2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(t)}-x^*||^2\leq(1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2+(2\gamma)(2\gamma L_b-1)\mathbb{E}(f(x^{(t-1)})-f(x^*))+2\gamma^2\sigma_b^*$

Since $\gamma<\frac{1}{2L_b}$, $2\gamma L_b-1<0$. We also know that $\mathbb{E}(f(x^{(t-1)})-f(x^*))>0$ Hence

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2+(2\gamma)(2\gamma L_b-1)\mathbb{E}(f(x^{(t-1)})-f(x^*))+2\gamma^2\sigma_b^* \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2+2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(t-1)}-x^*||^2+2\gamma^2\sigma_b^*$

Let's build this inequality recursively and see how it unfolds:
$\mathbb{E}||x^{(1)}-x^*||^2 \leq (1-p\gamma)||x^{(0)}-x^*||^2+2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(2)}-x^*||^2 \leq (1-p\gamma)\mathbb{E}||x^{(1)}-x^*||^2+2\gamma^2\sigma_b^*$

$\mathbb{E}||x^{(2)}-x^*||^2 \leq (1-p\gamma)((1-p\gamma)||x^{(0)}-x^*||^2+2\gamma^2\sigma_b^*)+2\gamma^2\sigma_b^*$

We can see that $\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-\gamma p)^t||x^{(0)}-x^*||^2+\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^*$
Let's look at the term $\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^*$.
$\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^* < \sum_{n=0}^{\infty}(1-p\gamma)^n 2\gamma^2\sigma_b^* = \frac{1}{p\gamma}2\gamma^2\sigma_b^* = \frac{2\gamma\sigma_b^*}{p}$
Hence, we can see that
$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-\gamma p)^t||x^{(0)}-x^*||^2+\sum_{n=0}^{t-1}(1-p\gamma)^n 2\gamma^2\sigma_b^* \leq (1-\gamma p)^t||x^{(0)}-x^*||^2+\frac{2\gamma}{p}\sigma_b^*$
$\mathbb{E}||x^{(t)}-x^*||^2 \leq (1-\gamma p)^t||x^{(0)}-x^*||^2+\frac{2\gamma}{p}\sigma_b^*$

4 Algorithmic Stability of SGD

4.1 Stability of Randomized Iterative Algorithms

4.1.1 Generalization Error

Generalization error is a measure of how accurately a machine learning model can predict outcome values for previously unseen data. Specifically, it quantifies the difference in performance between training data and new, unseen data. A key goal in areas such as machine learning is to minimize this error, which indicates better model performance on new, unseen data. The generalization error of a learning algorithm can be formally defined as the difference between the expected loss over the distribution of all possible data and the empirical loss calculated on the training set. Following is a build-up for a formal definition of the generalization error adopted from [**Stability**],
Consider a supervised learning setting where:

- The samples are drawn from a $\mathcal{D}$ unknown distribution.
- The samples $S = (x_1, x_2, \ldots, x_n)$ is samples i.i.d from the aforementioned distribution.
- The main objective is to find the model w with an associated loss function $f(w; x)$
- n denotes the number of samples.

The algorithm tries to find the model w with a minimum population risk defined as the expected loss over $\mathcal{D}$,

$$R[w] \triangleq \mathbb{E}_{x \sim \mathcal{D}}[f(w; x)] \tag{208}$$

However, it is important to note that, not knowing $\mathcal{D}$ makes it impossible to measure $R[w]$ directly. However, having a sufficient number of samples allow us to estimate population risk through the empirical average of the loss function over the samples (empirical risk),

$$R_S[w] \triangleq \frac{1}{n} \sum_{i=1}^{n} f(w; x_i) \tag{209}$$

Now the generalization error of the model w is the difference between the population risk and the empirical risk,

$$\text{Generalization Error} = R[w] - R_s[w] \tag{210}$$

In many applications, the model parameters w are learned using the sample data and response variables. For example, in linear regression, the goal is to establish a linear relationship between the features (or predictors) and the response variable. This relationship is expressed in the form $y = w^T x + b$, where y is the response, x is the feature vector, w is the vector of weights, and b is a bias term. By minimizing the empirical risk, typically represented by the mean squared error between the predicted values and the actual values in the training data, the model learns the parameters that best fit the data.

To learn these parameters effectively, especially in complex or large-scale settings, randomized algorithms such as Stochastic Gradient Descent (SGD) are frequently employed. This makes it computationally more efficient than batch gradient descent, particularly with large datasets. Here the random nature of the algorithm develops through the random selection of data subsets. With this insight it can be seen that the model w can be taken as a function of data S through a randomized algorithm A, where, $w = A(S)$. This allows one to define an expected generalization error over the randomness of the samples and the algorithm.

$$\epsilon_{gen} \triangleq \mathbb{E}_{S,A}[R_S[A(S)] - R[A(s)]] \tag{211}$$

4.2 Stability of SGD

According to Hardt et al. (2016), the stability of a learning algorithm, such as Stochastic Gradient Descent (SGD), can directly influence the generalization error. An algorithm is considered stable if small changes in the training set result in small changes in the outcome. The stability of SGD is quantified in the context of how the learning rate and other parameters are adjusted during the training process.

4.3 Algorithmic Stability and Generalization

The paper establishes a connection between the stability of SGD and its generalization performance:

$$\text{Generalization Error of SGD} \leq O\left(\frac{1}{\sqrt{n}}\right) \tag{212}$$

This inequality shows that the generalization error of SGD decreases at a rate proportional to the inverse square root of the number of training samples, under certain conditions related to the learning rate and the smoothness of the loss function.

4.4 Conclusion

Understanding and minimizing generalization error is crucial for improving the performance of machine learning models on new, unseen data. The stability properties of algorithms like SGD play a significant role in their ability to generalize well.

5 Numerical SGD Experiments

5.1 SGD: Ridge Regression

For the first set of experiments, Stochastic Gradient Descent will be used to solve the Ridge Regression Problem.